PROBABILISTIC COMPARISON OF SURVIVAL ANALYSIS MODELS USING SIMULATION AND CANCER DATA

YONG XU\textsuperscript{1} AND CHRIS P. TSOKOS\textsuperscript{2}

\textsuperscript{1}Department of Mathematics and Statistics, University of South Florida
Tampa, FL 33620 USA
\textit{E-mail:} yxu2@mail.usf.edu

\textsuperscript{2}Department of Mathematics and Statistics, University of South Florida
Tampa, FL 33620 USA
\textit{E-mail:} profcpt@cas.usf.edu

\textbf{ABSTRACT:} The object of the present study is to probabilistically evaluate commonly used methods to perform survival analysis of medical patients. Our study includes evaluation of parametric, semi-parametric and nonparametric analysis of probability survival models. We will evaluate the popular Kaplan-Meier (KM), the Cox Proportional Hazard (Cox PH), and Kernel density (KD) models using both Monte Carlo simulation and using actual breast cancer data. The first part of the evaluation will be based on how these methods measure up to parametric analysis and the second part using actual cancer data. As expected, the parametric survival analysis when applicable gives the best results followed by the not commonly used nonparametric Kernel density approach for both evaluations using simulation and actual cancer data.

\textbf{AMS (MOS) Subject Classification.} 62N01, 62N02 and 62N05.

\section{1. INTRODUCTION}

Wikipedia defines survival analysis as a branch of statistics which deals with death in biological organisms and failure in mechanical systems. Scientists have developed and used many different probabilistic survival analysis methods including parametric, nonparametric and semi-parametric models. In the present study we will compare all commonly used methods and propose which ones give the best probabilistic survival results.

The first part of our study is based on simulating data from a well defined probability failure distribution by identifying the sample size so that the maximum likelihood estimates converge to the assumed parametric values in a Monte Carlo simulation procedure. Using this information we develop and compare the parametric estimated probabilistic survival function with the Kernel density (nonparametric), and the popular Kaplan-Meier (KM) model.
The second part of our study uses actual survival time of breast cancer data to compare the above mentioned survival models, in addition to the Cox Proportional (Cox PH) survival hazard function.

Upon completing the evaluation, we will propose a ranking of the analytical methods evaluated for performing survival analysis. The breast cancer data that we used was given by N. A Ibrahim where the analysis and results were published at [1] and [14]. Some other relevant references that we use in the present study are [2], [11], [12], [13], [15], [18] and [19].

2. SURVIVAL ANALYSIS USING SIMULATION

2.1 PARAMETRIC SURVIVAL ANALYSIS

For our parametric Monte Carlo simulation, we shall assume that the failure data is being probabilistically characterized by the two parameter gamma probability density function (pdf), given by

$$f(x; \alpha, \beta) = \begin{cases} \frac{x^\alpha e^{-x}}{\beta^\alpha \Gamma(\alpha)}, & \alpha > 0, \beta > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and } x \geq 0$$

(2.1)

where $\alpha$ and $\beta$ are the shape and scale parameter, respectively and $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$.

The cumulative distribution function (CDF) of 2.1 is given by

$$F(x; \alpha, \beta) = \frac{\Gamma_x(k, \frac{x}{\beta})}{\Gamma(k)}; \quad x \geq 0$$

(2.2)

where $\Gamma_x(\alpha, x) = \int_0^x t^{\alpha-1}e^{-t}dt$ is the lower incomplete gamma function.

The survival function of the gamma pdf is given by

$$S(x; \alpha, \beta) = 1 - \frac{\Gamma_x(k, \frac{x}{\beta})}{\Gamma(x)}; \quad \alpha > 0, \beta > 0, x \geq 0$$

(2.3)

and the hazard function is of the form

$$h(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha) - \Gamma_x(\alpha)}; \quad x \geq 0, \alpha > 0$$

(2.4)

Also, the cumulative hazard function is given by

$$H(x) = -\log(1 - \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)}); \quad x \geq 0, \alpha > 0$$

(2.5)

where $\Gamma_x(\alpha)$ is the incomplete gamma function.

For the first part of our study we assume that $\alpha = 3.0$, $\beta = 2.0$ and rate parameter $\lambda = 0.5$. We simulated a sample of $n=300$ failure time where the maximum likelihood estimates resulted in $\hat{\alpha} = 2.984$, $\hat{\beta} = 2.903$ and $\hat{\lambda} = \frac{1}{\beta} = 0.4977$ which closely converge to the assumed true parameters. Thus, for $n=300$ failures, we have a very good random sample to begin parametrically our evaluation process.
Thus, the parametric true survival and hazard functions are given by

\[ S(t,3,2) = 1 - \frac{\Gamma_t(3, \frac{t}{2})}{\Gamma(3)}, \]

and the true hazard function is given by

\[ h(t) = \frac{t^2 e^{-t}}{\Gamma(3) - \Gamma_t(3)}, \quad t \geq 0, \alpha > 0. \]

The estimated parametric survival function is given by

\[ \hat{S}(t, \hat{\alpha}, \hat{\beta}) = 1 - \frac{\Gamma_t(2.984, \frac{t}{2.0093})}{\Gamma(2.984)}, \]

and the estimated parametric hazard function is given by

\[ \hat{h}(t) = \frac{t^{2.0093} e^{-t}}{\Gamma(2.984) - \Gamma_t(2.984)}, \]

respectively.

For comparison purposes, we shall refer to the survival functions, \( S(t, 3, 2) \), as the true parametric probabilistic survival curve and \( \hat{S}(t; 2.984, \hat{\gamma} = 0.4977) \), as the parametric estimates. Figure 2.1, below, gives a graphical display of the two probabilistic survival curves and Figure 2.2 below gives the corresponding estimate of the hazard function, \( \hat{H}(t) \).

![Figure 2.1 Survival plots for true and fitted parametric analysis](image)

It is clear that both estimated plots and true plots are almost identical as a function of age.
Figure 2.2 Cumulative hazard plot for true and fitted parametric analysis

**Probability residual analysis**

In order to clearly show the definition of the probability residuals, we made the Figure 2.3 below from a random gamma distribution. We digitized the time into n=300 single point. At each single point we define the difference between the two survival curves as our probability residuals. \( \gamma_i = \hat{S}_{true}(t_i) - \hat{S}_{fitted}(t_i) \) for \( i=1,2,3,\ldots,n \).

Figure 2.3 Illustration of the definition of probability residuals
Then we proceed to calculate the mean probability residual, the sample variance, the sample standard deviation, and sample standard error. Thus, the mean of the probability residual is 0.000655, sample variance is 2.675e-07, sample standard deviation is 0.00052 and standard error is 2.986e-05. These numbers attest to the quality of the suggest model.

2.2 Kaplan-Meier Method

The Kaplan-Meier method is the most popular in developing the survival functions for a given set of failure times. Some relevant references on the subject matter are [3], [4], [5], [6], [7] and [8]. The survival function, \( S(t) \), is the probability that an item from a given population will have a survival time exceeding \( t \). Let us consider a random sample of size \( n \) of the failure observed times until death, that is, \( t_1, t_2, \ldots, t_n \) and arranging them in the following manner

\[
t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_{n-1} \leq t_n.
\]

Define \( n_i \) as the number of patients at risk just prior to time \( t_i \) and let \( d_i \) be the number of deaths at exactly time \( t_i \).

**Analytical Form**

The estimate of the survival function of the Kaplan-Meier model is given by

\[
\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}.
\]

(2.10)

The estimate of the cumulative hazard function is given by

\[
\hat{H}(t) = -\ln\left( \prod_{t_i < t} \frac{n_i - d_i}{n_i} \right).
\]

(2.11)

For the Monte Carlo simulation of \( n=300 \) failure times we have the plots in Figure 2.4, the true parametric form of \( S(t) \) and with the Kaplan-Meier estimate of the survival curve along with the 95% confidence limits.

Figure 2.5, below, displays the estimated hazard function of the KM model along with the true parametric plot with 95% confidence limits.

**Probability Residual Analysis**

Consider the difference between true parametric survival curve and the estimate \( S(t) \) as our probability residual as defined before. Thus, the mean of the probability residual is -0.00345, sample variance is 0.000267, sample standard deviation is 0.01635304 and standard error is 0.000944143. Clearly as expected the parametric survival models gives continuously better result than the Kaplan-Meier model.

2.3 Kernel Density Estimation
A very powerful nonparametric method that has been used extensively to estimate the probability density of a certain data that does not follow any well known classical pdf is the Kernel Density estimation. For more information on this subject method see [11],[12],[13],[18] and [19].

For a given set of data $x_1, x_2, \ldots, x_n$ independent and identically distributed sample of a random variable, then the kernel density estimate of the probability density function is given by
\[ \hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \]  

(2.12)

and K is the kernel and h is the optimal bandwidth.

Given below, Table 2.1 is a list of the most commonly used kernels.

<table>
<thead>
<tr>
<th>Name of kernel</th>
<th>Math form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>( K(u) = \frac{1}{2}1_{(</td>
</tr>
<tr>
<td>Triangle</td>
<td>( K(u) = (1 -</td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>( K(u) = \frac{3}{4}(1 - u^2)1_{(</td>
</tr>
<tr>
<td>Quartic</td>
<td>( K(u) = \frac{3}{4}(1 - u^2)^21_{(</td>
</tr>
<tr>
<td>Triweight</td>
<td>( K(u) = \frac{35}{32}(1 - u^2)^31_{(</td>
</tr>
<tr>
<td>Gaussian</td>
<td>( K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} )</td>
</tr>
<tr>
<td>Cosine</td>
<td>( K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)1_{(</td>
</tr>
</tbody>
</table>

The most frequently used optimal bandwidth is shown in Table 2.2, below.

<table>
<thead>
<tr>
<th>Name of optimal bandwidth</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal 0 (Nrd0)</td>
<td>Represents the bandwidth of a Gaussian kernel. The numerical value is 0.9<em>min[standard deviation(SD), Interquartilerange (IQR)] / (1.34</em>(sample size)^(-1/5))</td>
</tr>
<tr>
<td>Normal (nrd)</td>
<td>The numerical value is 1.06<em>min(SD, IQR) / (1.34</em>(sample size)^(-1/5))</td>
</tr>
<tr>
<td>unbiased cross-validation (ucv)</td>
<td>For unbiased cross-validation</td>
</tr>
<tr>
<td>biased cross-validation (bcv)</td>
<td>For biased cross-validation</td>
</tr>
<tr>
<td>Select (SJ)</td>
<td>select the bandwidth using pilot estimation of derivatives</td>
</tr>
</tbody>
</table>

**Analytical Form**

We run the models for all the combinations of different kernels and optimal bandwidth. The Epanechnikov kernel and the proposed optimal bandwidth gave the best results.
The Kernel Density survival function is given by

\[ S_h(x) = 1 - \sum_{X \leq x} \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right), \]  

(2.13)

where \( K \) is the kernel and \( h \) is the optimal bandwidth, respectively.

The hazard function of Kernel Density \( S_h(x) \) method is given by

\[ h_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right), \]

with

\[ h = \frac{1.06\min(SD(x), IQR(x))}{1.34n^{-\frac{1}{5}}}, \]  

(2.14)

Thus the estimated value of \( S_h(x) \) is given by

\[ \hat{S}_h(x) = 1 - \sum_{X \leq x} \frac{1}{N^\frac{3}{4} 1.34n^{-\frac{1}{5}}} \sum_{i=1}^{N} 3^\frac{3}{4} (1 - \left(\frac{x-x_i}{1.06\min(SD(x), IQR(x))}\right)^2), \]

(2.16)

and the estimate of the hazard function is given by

\[ \hat{h}_h(x) = \frac{\sum_{i=1}^{N} 3^\frac{3}{4} (1 - \left(\frac{x-x_i}{1.06\min(SD(x), IQR(x))}\right)^2)}{1 - \sum_{X \leq x} \frac{1}{N^\frac{3}{4} 1.34n^{-\frac{1}{5}}} \sum_{i=1}^{N} 3^\frac{3}{4} (1 - \left(\frac{x-x_i}{1.06\min(SD(x), IQR(x))}\right)^2)}. \]

(2.17)

Thus, the estimate of the survival function, \( \hat{S}_h(x) \) for the KD method is given by Figure 2.5, below. It is clear that the KD approach to \( \hat{S}(t) \) is almost identical to the parametric survival function.

The estimates of the hazard function of the true and KD estimates are shown by Figure 2.6, below. Thus, it is clear from the graph, Figure 2.7, that the true \( \hat{H}(t) \) and the Kernel density estimates are approximately the same.

**Probability Residual Analysis**

Consider the difference between the true and parametric survival curve as our probability residual. Then the sample mean of the probability residual is -0.0023, sample variance is 0.0001, sample standard deviation is 0.0163 and standard error is 0.000579.

Table 2.3 below gives us the summary and comparison of the three probability residuals.
From the Table 2.3 we can see that the parametric model as expected is much better than other two. KM and KD are very close using the mean of the probability residuals but KD’s standard deviation and standard error is about 50% smaller.
Table 2.3 Residual analysis

<table>
<thead>
<tr>
<th>Model Comparison</th>
<th>Mean</th>
<th>Variance</th>
<th>SD</th>
<th>SE</th>
<th>Rank of Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted parametric vs. True parametric</td>
<td>0.00065</td>
<td>2.67e-07</td>
<td>0.00052</td>
<td>2.986e-05</td>
<td>1</td>
</tr>
<tr>
<td>Kaplan-Meier vs. True parametric</td>
<td>-0.0034</td>
<td>0.00027</td>
<td>0.0163</td>
<td>0.000944</td>
<td>3</td>
</tr>
<tr>
<td>Kernel density vs. True parametric</td>
<td>-0.0023</td>
<td>0.0001</td>
<td>0.01</td>
<td>0.000579</td>
<td>2</td>
</tr>
</tbody>
</table>

than the popular KM model. Therefore if parametric analysis is not justified, we recommend that the KD should be used for developing the survival model.

From this Figure 2.8 we can observe that KD is much smoother than KM and is closer to true parametric in the middle but in the beginning it tends to be too conservative.

3. STATISTICAL MODEL VALIDATION FOR NONCENSORED CANCER DATA

3.1 Parametric Data Validation Analysis
It is well known that parametric analysis will result in the best estimate of the probability survival curves. However, if we cannot justify parametric analysis, then we must proceed with nonparametric estimates using KD, KM or semi-parametric Cox PH survival models.

For the breast cancer data, [14], we have survival times of 641 breast cancer patients with 48 being uncensored. Proceeding with goodness of fit statistical methods, using the Kolmogorov-Smirnov test we have identified that the two parameter gamma probability distribution fits the breast cancer data quite well. The gamma probability failure distribution with the maximum likelihood estimates of the shape and scale parameter is given by equation 3.1, below,

\[
\hat{f}(t; \hat{\alpha}, \hat{\beta}) = \begin{cases} 
\frac{t^{\hat{\alpha}-1}e^{-t/\hat{\beta}}}{\Gamma(\hat{\alpha})}, & t \geq 0 \\
0, & \text{otherwise}
\end{cases}
\]

A graphical form of the gamma survival model is given below by Figure 3.1.

![Figure 3.1 Gamma fitted probability density function for the cancer data](image)

Figure 3.1 Gamma fitted probability density function for the cancer data

Thus, we can proceed to obtain the parametric survival function and proceed to develop and use it as reference to compare it with the KD, KM and Cox PH survival models, [16], [17].
The estimated parametric survival function based on the gamma pdf is given by

\[ \hat{S}_p(t, \hat{\alpha}, \hat{\beta}) = 1 - \frac{\Gamma_t(1.997, t \frac{1.003}{t})}{\Gamma(1.997)}, \] (3.2)

and its corresponding estimated of the cumulative hazard function is given by

\[ \hat{H}_p(t, \hat{\alpha}) = -\log\left(1 - \frac{\Gamma_x(1.997)}{\Gamma(1.997)}\right). \] (3.3)

A graphical presentation of \( \hat{S}_p(t) \) is shown below by Figure 3.2.

![parametric survival plot](image)

Figure 3.2 Plot of parametric survival model

### 3.2 Kaplan-Meier survival analysis

The estimate of the survival function of the Kaplan-Meier survival model is given by

\[ \hat{S}_{KM}(t) = \prod_{t_i < t} \frac{ni - di}{ni}. \] (3.4)

and the corresponding estimate of the cumulative hazard function is given by

\[ \hat{H}_{KM}(t) = -\ln\left(\prod_{t_i < t} \frac{ni - di}{ni}\right), \] (3.5)

where \( ni \) is the number of patients at risk just prior to time \( t_i \) and \( di \) is the number of deaths at exactly time \( t_i \).
Figure 3.3 Survival curve for parametric and KM models with 95% C.I.

Figure 3.4 the cumulative hazard plot for true and KM method

A graphical display of the estimated KM survival curve and estimated KM cumulative hazard curve along with 95% confidence limits is shown by Figures 3.3 and 3.4.

From the Figure 3.3 and 3.4, above, we can observe that the KM method is close to the parametric plot; however, since it is a step like function, one will prefer the parametric estimate which is a smooth curve.

Probability residual analysis

Since the probability residual analysis follows the same procedure discussed in the previous section, we can conclude that the mean of the probability residuals...
is 0.00949, sample variance is .0196 with sample standard deviation is 0.14014 and
standard error is 0.02023. Thus, the KM method is quite close to the parametric
method.

3.3 KD survival analysis

The estimated value of the KD survival function is given by

$$\hat{S}_{KD}(x) = 1 - \sum_{X \leq x} \frac{1}{N \bar{X}_{i,0.06 \min(SD(x), IQR(x))}} \sum_{i=1}^{N} \frac{3}{4} \left(1 - \left(\frac{x - x_i}{1.34n^{-\frac{1}{2}} \bar{X}_{i,0.06 \min(SD(x), IQR(x))}}\right)^2\right), \quad (3.6)$$

and the estimate of the hazard function is given by

$$\hat{h}_{KD}(x) = \frac{\sum_{X \leq x} \frac{1}{N \bar{X}_{i,0.06 \min(SD(x), IQR(x))}} \sum_{i=1}^{N} \frac{3}{4} \left(1 - \left(\frac{x - x_i}{1.34n^{-\frac{1}{2}} \bar{X}_{i,0.06 \min(SD(x), IQR(x))}}\right)^2\right)}{1 - \sum_{X \leq x} \frac{1}{N \bar{X}_{i,0.06 \min(SD(x), IQR(x))}} \sum_{i=1}^{N} \frac{3}{4} \left(1 - \left(\frac{x - x_i}{1.34n^{-\frac{1}{2}} \bar{X}_{i,0.06 \min(SD(x), IQR(x))}}\right)^2\right)}. \quad (3.7)$$

Figure 3.5 the true and KD survival curve

A graphical presentation of the estimated KD survival curve and estimated KD
cumulative hazard curve is given by Figures 3.5 and 3.6 along with the parametric
results.

From the Figures 3.5 and 3.6 we can observe that the KD curves for survival and
cumulative hazard are closer to the parametric curves, which indicates that the KD
survival method seems better than the KM survival method.

Probability residual analysis
The mean of the probability residual is 0.005529787, sample variance is 0.000553 with sample standard deviation is 0.02352216 and standard error is 0.003395132. Clearly, the KD method gives better estimates than the KM method in terms of both sample mean and sample standard error.

3.4 Cox PH model

The Cox PH model is a very popular semi-parametric method that has been used extensively to estimate the survival probability function of a given set of data that characterizes the failure time of a given patient. For more information on this subject see [3],[5],[6],[10], [20] and [21].

**Analytical form**

Hazard function of the Cox PH model is given by

$$h_t(x) = h_0 \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}),$$

and its survival function is given by

$$S_t(x) = \exp(- \int_{-\infty}^{\infty} h_0 \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}) dt),$$

where $h_0$ is the baseline hazard and all betas are the coefficients of the covariates.

The data set contains these covariates as shown in Table 3.1, below.

From table 3.1, we start with covariate of the first order, namely, stnum, tx, pathsize, hist, hrlevel, hgb, nodediss, age and all possible 2nd terms and interactions. By using a stepwise selection with minimum Akaike’s information criterion (AIC), we
Table 3.1 The data set variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>stnum</td>
<td>Patient ID</td>
</tr>
<tr>
<td>tx</td>
<td>Randomized treatment: T=tamoxifen, B=radiation + tamoxifen</td>
</tr>
<tr>
<td>pathsize</td>
<td>Size of the tumour in cm</td>
</tr>
<tr>
<td>hist</td>
<td>Histology: DUC=Ductal, LOB=Lobular, MED=Medullary, MIX=Mixed, OTH=Other</td>
</tr>
<tr>
<td>hrlevel</td>
<td>Hormone receptor level: NEG=Negative, POS=Positive</td>
</tr>
<tr>
<td>hgb</td>
<td>Haemoglobin in g/l</td>
</tr>
<tr>
<td>nodediss</td>
<td>Whether axillary node dissection was done: Y=Yes, N=No</td>
</tr>
<tr>
<td>age</td>
<td>Age in years</td>
</tr>
</tbody>
</table>

obtain the final model with only two first order terms that are significant, namely, tx and stnum.

Table 3.2 Cox model summary

|        | coef     | exp(coef) | se(coef) | z      | Pr(>|z|) |
|--------|----------|-----------|----------|--------|----------|
| txT    | -1.660047| 0.190130  | 0.763126 | -2.175 | 0.0296   |
| stnum  | 0.005599 | 1.005615  | 0.002348 | 2.384  | 0.0171   |

Furthermore, the quality of the selected model based on the three statistical criteria is given in the Table 3.3, below, that support the quality of the fitted model,

The fitted survival function is given by

$$\hat{S}_{Coxph}(x) = \exp\left(-\int_{-\infty}^{\infty} h_o \exp(-1.66x_{i1} + .005599x_{i2})dt\right), \quad (3.10)$$

with the corresponding estimate of the hazard function

$$\hat{H}_{Coxph}(x) = \int_{-\infty}^{\infty} h_o \exp(-1.66x_{i1} + .005599x_{i2})dt. \quad (3.11)$$

From Tables 3.2 and 3.3 we can conclude that Cox PH model does not seem to fit the data very well because the p-value of the Likelihood ratio test (LRT), Wald test and Score test are all greater than .05.

Table 3.3 Cox model’s significance

<table>
<thead>
<tr>
<th>test</th>
<th>value</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood ratio test</td>
<td>5.39</td>
<td>2</td>
<td>0.06746</td>
</tr>
<tr>
<td>Wald test</td>
<td>5.69</td>
<td>2</td>
<td>0.05822</td>
</tr>
<tr>
<td>Score (logrank) test</td>
<td>5.74</td>
<td>2</td>
<td>0.05684</td>
</tr>
</tbody>
</table>
From Figures 3.7 and 3.8 we can see that the Cox PH curve does not seem to fit the data very well.

![Survival plot for parametric and Cox PH model with 95% CI](image1.png)

**Figure 3.7** The survival plot for parametric and Cox PH model with 95% CI

![Cumulative hazard plot for parametric and Cox PH model](image2.png)

**Figure 3.8** The cumulative hazard plot for parametric and Cox PH model

From Table 3.4 we can conclude that if we exhaust all possible parametric choices, then we have to perform nonparametric analysis. In this situation we will recommend the KD method in terms of its smoothness and less standard error from the probability residuals analysis.

A graphical comparison of survival plots for KD, parametric, KM and Cox PH model are summarized by Figure 3.9, below,
Figure 3.9 The survival curves for KD, parametric, KM and Cox PH model

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Rank of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox PH vs. fitted parametric</td>
<td>0.0095</td>
<td>0.1401</td>
<td>0.02023</td>
<td>3</td>
</tr>
<tr>
<td>Kaplan-Meier vs. fitted parametric</td>
<td>0.0095</td>
<td>0.1401</td>
<td>0.0202</td>
<td>2</td>
</tr>
<tr>
<td>Kernel density vs. fitted parametric</td>
<td>0.00553</td>
<td>0.0235</td>
<td>0.003395</td>
<td>1</td>
</tr>
</tbody>
</table>

From a visual evaluation of the four survival curves, we can conclude that the KD model comes the closest to the parametric model followed by KM and Cox PH. Thus, using the real cancer data we can conclude that the KD is better than the KM and Cox PH models. This is consistent with MC simulation that we initially performed. It is recommended that when the censoring rate is small then KD is the best model to use.

4. STATISTICAL MODEL VALIDATION FOR CENSORED DATA

Quite often we deal with censored data due to limited and difficult experimental conditions. In the present study we are interested in investigating how KD analysis performs under a censored data situation. The problem is that we will never know...
the true state of nature under the censored circumstances and the only information we are certain of is that by the time it is censored the patient is still alive.

How to conduct a goodness of fit test for censored data is still an open problem. Edel A. Pena, [9], discussing the subject matter stated that we can only reject some probability distributions but still can not have the unique best distribution to probabilistically characterize the censored data. In this study we will perform KM, KD, Cox PH and parametric survival analysis for censored data and evaluate their response.

4.1 Parametric Survival Analysis

Despite the difficulties of the censored data, we find the best possible fit for the cancer data follows two parameter Weibull distribution.

The two parameter Weibull pdf is given by 4.1, below,

\[
f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}, & \alpha > 0, \quad \beta > 0 \quad \text{and} \quad x \geq 0 \\ 0, & \text{otherwise} \end{cases}
\]  

(4.1)

where \( \alpha \) and \( \beta \) are the shape and scale parameter respectively.

The cumulative distribution function (CDF) of 4.1 is given by

\[
F(x; \alpha, \beta) = 1 - e^{-\left(\frac{x}{\beta}\right)^\alpha}, \quad \alpha > 0, \quad \beta > 0 \quad \text{and} \quad x \geq 0
\]

(4.2)

The survival function of the Weibull distribution is given by

\[
S(x; \alpha, \beta) = e^{-\left(\frac{x}{\beta}\right)^\alpha}, \quad \alpha > 0, \quad \beta > 0 \quad \text{and} \quad x \geq 0
\]

(4.3)

and the hazard function is of the form

\[
h(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1}, \quad \alpha > 0, \quad \beta > 0 \quad \text{and} \quad x \geq 0.
\]

(4.4)

From the parametric analysis, we obtain the estimates \( \hat{\alpha} = 1.0962 \) and \( \hat{\beta} = 59.243 \).

The estimated parametric survival function is given by

\[
\hat{S}(t; \hat{\alpha}, \hat{\beta}) = e^{-\left(\frac{t}{59.243}\right)^{1.0962}}, \quad x \geq 0
\]

(4.5)

We should assume that this parametric survival model will be the best possible survival model for this censored cancer data. We will utilize this survival model to evaluate the performance of the other three nonparametric survival methods, namely, KM, Cox PH and KD survival methods.

4.2 KM Survival Analysis
The survival probability estimate of the censored data using the KM model with the parametric survival model is given by Figure 4.1, below. In comparing the KM survival curve with the parametric model using the probability residuals, we have found that the mean of the probability residual is 0.000495, sample variance is $4.596 \times 10^{-5}$, sample standard deviation is 0.0068 and standard error is 0.0024.

![Figure 4.1 The survival plot for KM model](image)

4.3 Cox PH model

To select the best possible Cox PH model for censored data, we consider the model has all terms significant with the minimum AIC. Through statistical testing we have found that six first order terms and two interactions significantly contribute to the response variable. These attributable variables are tx, pathsize, nodediss, age, hrlevel, stnum, tx:age and nodediss:hrlevel. Thus, for the subject data and the attributable variables using the Cox PH model we plot the probability survival curve with the parametric survival curve and they are shown by Figure 4.2, below. In comparing the Cox PH curve with the estimated parametric survival curve, we found the mean of the probability residual is 0.028, sample variance is 0.000347, sample standard deviation is 0.0186 and standard error is 0.0066.

4.4 KD survival analysis

Despite the difficulties in working with censored data, we proceeded to use the nonparametric KD procedure to estimate the survival curve together with the fitted parametric survival curve. The results are shown by Figure 4.3, which are different than what we have found using KM and Cox PH models in terms of the smoothness of the curve.
In comparing the KD survival curve with the fitted parametric model, we have found the mean of the probability residual is 0.00883, sample variance is $2.942 \times 10^{-5}$, sample standard deviation is 0.0054 and standard error is 0.00191.

Table 4.1, below, summarizes the response of the three survival analysis models, KM, Cox PH and KD, in comparison with the parametric model using the two parameter Weibull probability density function to characterize the failures. Thus, if we assume that we can proceed to statistically analyze the censored data, all three survival models performed well, but the edge goes to the KD model in terms of the smaller sample variance and standard error.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Rank of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox PH vs. fitted parametric</td>
<td>0.1944</td>
<td>0.274</td>
<td>0.01582</td>
<td>3</td>
</tr>
<tr>
<td>Kaplan-Meier vs. fitted parametric</td>
<td>0.1746</td>
<td>0.2768</td>
<td>0.01598</td>
<td>2</td>
</tr>
<tr>
<td>Kernel density vs. fitted parametric</td>
<td>-0.0164</td>
<td>0.3008</td>
<td>0.0434</td>
<td>1</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

The present study consists of three parts in comparing the effectiveness of three survival analysis models, namely, KM, Cox PH and KD. Initially, using Monte Carlo simulation we compare the subject models with parametric survival models and found that the proposed KD survival model gives as good results, if not better, than the KM.
The second part consists of using actual uncensored breast cancer data. Performing a similar evaluation, the results support that the proposed KD model gives results in better estimates than the popular KM and Cox PH models with interactions.

Thirdly, we performed the same analysis with actual censored breast cancer data. Although working with censored data is quite difficult to justify such an analysis, under the circumstances we analyzed the data and the results are similar to the Monte Carlo simulation and using the uncensored data.

ACKNOWLEDGMENTS

The authors wish to acknowledge N.A Ibrahim for supplying the original data also wish to express appreciation to Dr. James Kepner, Vice President of the American Cancer Society for the useful discussion during the progress of the present study.

REFERENCES


